Statistical Tests to Audit Investigative Stops

Ravi Shroff

Data-centric technologies are transforming criminal justice in the United States. These systems and techniques are usually designed to accomplish one or more of the following goals: to better collect criminal justice information via devices like body cameras or acoustic gunshot sensors; to audit the effectiveness and constitutionality of historical practices; and to improve decision-making processes like pretrial detention determinations by judges or officer deployments by police departments. These uses of "big data" and associated algorithmic procedures offer the promise of increased accountability, efficiency, and equitability. They also raise legal concerns that are inadequately addressed by existing jurisprudence, and policy concerns associated with current and planned criminal justice processes. This paper examines these concerns in the context of the investigative stop, during which a police officer briefly detains a criminal suspect—and may perform a limited search, or frisk, of the suspect's outer garments—and statistical tests for racial discrimination in programs of these stops.[1] One of these tests is described in detail, and then two broad questions are raised regarding how statistical tests for racial discrimination can be implemented and supported in practice.[2]

The main Supreme Court ruling concerning the police stop is *Terry v. Ohio*, which clarified that a standard of "reasonable suspicion" of criminality is required for a stop to be made, and that reasonable suspicion that the suspect is armed and dangerous is additionally required for the frisk.[3] Although extensively used,[4]

[1] This article considers both pedestrian and vehicle stops, and will refer to them interchangeably as investigative stops, *Terry* stops, or the more colloquial "stop-and-frisk."

[2] For a more comprehensive treatments of these ideas, *see* Sharad Goel, Justin M. Rao & Ravi Shroff, *Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy*, 10 ANNALS APPLIED STAT. 365 (2016) [hereinafter Goel et al., *Precinct or Prejudice?*]; Sharad Goel, Maya Perelman, Ravi Shroff & David Alan Sklansky, *Combatting Police Discrimination in the Age of Big Data*, 20 NEW CRIM. L. REV. 181 (2017) [hereinafter Goel et al., *Combatting Police Discrimination*].

[3] Terry v. Ohio, 392 U.S. 1, 30 (1968).

[4] Between 2014 and 2015, the Chicago Police Department (CPD) carried out over 1.3 million stops, ARLANDER KEYS, THE CONSULTANT'S FIRST SEMIANNUAL REPORT ON THE INVESTIGATORY STOP AND PROTECTIVE PAT DOWN AGREEMENT FOR THE PERIOD JANUARY 1, 2016 – JUNE 30, 2016, at 20–21 (2017), and between 2008 and 2012, the New York City Police Department (NYPD) carried out almost 3 million stops, *Stop-and-Frisk Data*, N.Y. CIVIL LIBERTIES UNION, https://www.nyclu.org/en/stop-and-frisk-data [https://perma.cc/63ZQ-RBHE]. More recently, the

programs of police stops remain controversial. Proponents claim that stop-and-frisk is an effective tool in interrupting and deterring criminal activity, because it is not subject to the more stringent standard of probable cause that governs arrests and searches.[5] Critics, however, argue that programs of police stops impose burdens on those detained, and do not efficiently accomplish the goal of getting weapons and drugs off the streets.[6] Moreover, they argue that stops are frequently conducted without reasonable suspicion, in violation of the Fourth Amendment, and that stop-and-frisk policies have been implemented in a racially discriminatory manner, in violation of the Equal Protection Clause of the Fourteenth Amendment.[7] Although this paper focuses on Equal Protection violations, it should be noted that general crime deterrence by itself is not an adequate legal justification for performing a *Terry* stop, and that the evidence for deterrence effects of stop-and-frisk is mixed.[8]

A number of powerful statistical tests have been used within the last decade to determine the presence of racial bias in *Terry* stops (these tests can also be used to test for bias toward other protected classes). In a *benchmark* test, the racial distribution of stopped *Terry* suspects is compared to a known benchmark distribution, such as the racial composition of the local residential population, or of local arrestees for violent crimes the previous year. Differences between the distribution of *Terry* suspects and the benchmark can suggest racial bias.[9] However, the plausibility of these benchmark tests often depends on strong assumptions about the benchmark distribution, e.g., that the racial composition of the local residential population mirrors the racial composition of those exhibiting reasonable suspicion sufficient to justify a *Terry* stop. Another class of *outcome* tests compares contraband recovery rates by race. If a lower proportion of black stops motivated by suspected weapon possession in fact recover a weapon—compared to the proportion of similarly motivated white stops that recover a weapon—this suggests that a lower standard of evidence was used by

NYPD has dramatically curtailed the use of stop-and-frisk; fewer than 13,000 stops were conducted city-wide in 2016. *See* N.Y. CIVIL LIBERTIES UNION, *supra*.

[5] *See, e.g.*, Sam Roberts, *Raymond Kelly, Ex-Police Commissioner, Blames Mayor de Blasio for Rise in Killings*, N.Y. TIMES (Sept. 7, 2015), https://www.nytimes.com/2015/09/08/nyregion/raymond-kelly-ex-police-commissioner-blames-mayor-de-blasio-for-rise-in-killings.html [https://perma.cc/L33Z-B23B].

[6] *See, e.g.*, N.Y. CIVIL LIBERTIES UNION, *supra* note 4.

[7] *See id.*

[8] *See generally* Aaron Chalfin & Justin McCrary, *Criminal Deterrence: A Review of the Literature*, 55 J. ECON. LITERATURE 5 (2017).

[9] Andrew Gelman, Jeffrey Fagan & Alex Kiss, *An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias*, 102 J. AM. STAT. ASS'N 813, 816–20 (2007); Floyd v. City of New York, 959 F. Supp. 2d 540, 583–89 (S.D.N.Y. 2013).

officers when stopping black suspects versus white suspects.[10] Other tests compare the racial distribution of drivers stopped after dark (when race is hard to observe) to drivers stopped during daylight hours,[11] or use information about the race of the officer conducting the stop.[12]

A type of outcome test sometimes referred to as "stop-level hit rate" (SHR) analysis can be used to assess discrimination in stops.[13] A SHR test uses a statistical model to measure the degree of suspicion motivating a stop. Specifically, the model uses all recorded information available to an officer at the time a stop is made to estimate the chance that the stop will be "successful." This estimate is the stop-level hit rate. Since stops must be motivated by reasonable and articulable suspicion of criminal activity, "success" might mean that a stop of an individual suspected to be carrying a weapon in fact recovers a weapon.[14] In the study conducted in *Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy*, the authors applied SHR analysis to New York City's stop-and-frisk data by fitting a logistic regression model to the half-million stops from 2008 to 2010 that were motivated by suspected criminal possession of a weapon (CPW).[15] This model included variables measuring recorded values of, among other things, the suspect's demographics (e.g., age and gender), reasons cited by the officer for the stop (e.g., "suspicious bulge"), and location information (e.g., police precinct), and estimated the likelihood that the

[10] Outcome tests suffer from known issues with *infra-marginality* and *subgroup validity*. Ian Ayres, *Outcome Tests of Racial Disparities in Police Practices*, 4 JUST. RES. & POL'Y 131, 135–41 (2002). However, a more powerful "threshold test" has recently been devised that circumvents these issues; this threshold test was applied to stop-and-frisk data from New York City, and found evidence of bias. *See* Camelia Simoiu, Sam Corbett-Davies & Sharad Goel, *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANNALS APPLIED STAT. 1193 (2017); Emma Pierson, Sam Corbett-Davies & Sharad Goel, *Fast Threshold Tests for Detecting Discrimination* (Stanford Univ., Working Paper arXiv:1702.08536v2 [stat.ML], 2018).

[11] *See, e.g.*, Jeffrey Grogger & Greg Ridgeway, *Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness*, 101 J. AM. STAT. ASS'N 878 (2006); Emma Pierson et al., *A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States* (Stanford Univ., Working Paper arXiv:1706.05678v1 [stat.AP], 2017).

[12] *See, e.g.*, Kate Antonovics & Brian G. Knight, *A New Look at Racial Profiling: Evidence from the Boston Police Department*, 91 REV. ECON. & STAT. 163 (2009); Shamena Anwar & Hanming Fang, *An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence*, 96 AM. ECON. REV. 127 (2006).

[13] The numerical estimates provided by SHR analysis can also inform assessments of Fourth Amendment violations and improve *Terry* stop policies. *See* Goel et al., *Combatting Police Discrimination*, *supra* note 2, at 211–20.

[14] Success could also mean that a stop predicated on the suspicion of drug possession in fact recovers drugs, or that any stop results in an arrest. However, the validity of these tests can suffer when the outcome is subject to human bias (e.g., there is arguably more officer discretion involved in the outcome of arrest, compared to the outcome of weapon recovery).

[15] Goel et al., *Precinct or Prejudice?*, *supra* note 2, at 368–74.

stop would uncover a weapon. The model was found to be accurate, gave calibrated estimates, and was robust to reasonable errors in the measurement of values like age, weight, and height.

Given stop-level hit rates, one can then analyze how they vary by the suspect's race. In the analysis cited above, although approximately 50% of black CPW stops had an estimated hit rate of less than 1%, only approximately 20% of white CPW stops had a similarly low hit rate.[16] Thus, these low hit rate stops disproportionately affected black individuals compared to white individuals. Moreover, SHR tests can be used to understand whether differences in race-specific hit rate distributions arise for race-neutral reasons. For example, if officers have a lower threshold of suspicion (which is what a stop-level hit rate is estimating) for conducting stops in high crime areas, and high crime areas happen to be areas with a high proportion of minorities, this might explain differences in race-specific hit rate distributions absent discriminatory decisions by individual officers. However, when hit rates by race were examined *within* each location, disparities persisted.[17] Therefore, SHR analysis can yield evidence to reject non-discriminatory explanations for racial disparities, and provide support for the claim that stop-and-frisk in New York City was discriminatory toward black individuals. Stop-level hit rates can also inform policy by identifying types of stops that are both unlikely to recover contraband, and which also impose a disproportionate burden on a protected class.

The accuracy and value of statistical tests for discrimination in programs of *Terry* stops depends on the existence of detailed and comprehensive data that records important aspects of these stops. Fortunately, many police departments are now systematically collecting such data,[18] either as a result of settlements or injunctions, or a desire to improve their practices and accountability. Furthermore, these data are now being organized and made available through additional efforts by nonprofit organizations like the New York Civil Liberties Union,[19] or academic collaborations like the Stanford Open Policing Project.[20] However, important characteristics of stops are unavailable in many jurisdictions: only twelve of fifty states have provided the Stanford Open Policing Project with the required

[16] *Id.* at 375.

[17] *Id.* at 377–78.

[18] For example, the NYPD annually releases an electronic dataset of stops. *Stop, Question and Frisk Data*, N.Y. POLICE DEP'T, http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page [https://perma.cc/T4F2-TMKU] (last visited Mar. 10, 2018).

[19] *See* N.Y. CIVIL LIBERTIES UNION, STOP-AND-FRISK 2011: NYCLU BRIEFING (2012), https://www.nyclu.org/sites/default/files/publications/NYCLU_2011_Stop-and-Frisk_Report.pdf [https://perma.cc/LX4G-X8LE].

[20] *See* STANFORD OPEN POLICING PROJECT, https://openpolicing.stanford.edu/ [https://perma.cc/9MSP-8DC9] (last visited Mar. 10, 2018); *see also* Pierson et al., *supra* note 11.

information to conduct outcome tests for search decisions (as of September 2017).[21] Moreover, characteristics of stops that are available in data are recorded in many different ways. Chicago's investigatory stop report (ISR) form requires the officer to record the stop reason using seven checkboxes (e.g., "actions indicative of engaging in drug transaction") and a free-text narrative section, whereas New York's analogous UF-250 form provides 20 checkboxes (e.g., "proximity to crime location") and a description of the suspected crime to describe the stop reason.[22] Also, until recently, the CPD only collected information on stops which did not result in further enforcement action, whereas the NYPD—in theory, at least—collects information on every stop that an officer conducts.[23] These inconsistencies across cities limit the utility of the statistical methods described above in making clear and meaningful inferences about the presence of racial bias in stop-and-frisk.

Beyond data issues, it has historically been challenging to use statistical tests to support a legal argument demonstrating Equal Protection violations in criminal justice contexts. These claims are generally difficult to make because of the "two-pronged test" introduced in *Washington v. Davis*,[24] requiring plaintiffs to demonstrate both the disparate impact of an action on a protected class,[25] as well as discriminatory intent, i.e., differences in treatment intentionally directed at members of that protected class. While statistical tests have been used to show disparate impact, finding evidence of discriminatory intent is much more difficult. One reason for this is the Supreme Court ruling in *McClesky v. Kemp*,[26] where statistical evidence showing that blacks who killed whites were more than seven times as likely to be sentenced to death as whites who killed blacks was not considered "stark" enough to infer a finding of discriminatory intent in the

[21] *See* STANFORD OPEN POLICING PROJECT, *supra* note 20.

[22] *See* Chicago Police Dep't, Investigatory Stop Report CPD-11.910 (Rev. 7/17); N.Y. Police Dep't, Stop, Question and Frisk Report Worksheet PD344-151A (Rev. 11-02).

[23] The CPD's ISR form has recently changed, and the NYPD's UF-250 may also be amended to change the type of data collected and the format in which data are recorded. *See* Rocco Parascandola, *Stop-and-Frisk Forms Would Require More Details from NYPD*, N.Y. DAILY NEWS (Mar. 23, 2016, 1:47 PM), http://www.nydailynews.com/new-york/stop-and-frisk-forms-require-details-nypd-article-1.2574886 [https://perma.cc/D7WF-PZ6Z]; ACLU OF ILL., MARCH 2017 STOP & FRISK REPORT (2017), https://www.aclu-il.org/en/publications/march-2017-stop-frisk-report [https://perma.cc/3BFV-UF7B] (providing a comprehensive report on the CPD's stop practices).

[24] Washington v. Davis, 426 U.S. 229 (1976).

[25] In the context of *Terry* stops, this could mean showing that similarly situated defendants of a different race were not stopped, or that similarly situated stopped defendants of a different race were not frisked.

[26] McCleskey v. Kemp, 481 U.S. 279 (1987).

particular sentence of the defendant.[27] Lower courts have also been hostile to the use of statistical evidence from outcome tests to infer discriminatory intent, and have rejected benchmark tests because of the use of unconvincing benchmarks.[28] However, these tests were often rejected due to data quality and comprehensiveness issues, rather than problems with the theoretical foundations of the tests themselves. Note that from a policy perspective, disparate impact is important beyond its use in inferring discriminatory intent. It is bad policy to place extra burdens on historically disadvantaged communities, although this may be unavoidable depending on circumstance. Also, in the seminal stop-and-frisk case, *Floyd v. City of New York*, a federal judge used some statistical evidence to support her finding of a violation of the plaintiff's Fourteenth Amendment rights (in particular, a benchmark test, to show disparate impact), but still relied upon explicit statements made by NYPD employees in verifying discriminatory intent.[29]

Given the current and historical difficulties with using statistical methods to provide evidence for discrimination in programs of investigative stops, I raise and briefly discuss two questions aimed at addressing these issues.

First, ***how can the use of statistical tests for discrimination be effectively standardized and made routine to audit and improve police actions?*** One obstacle to this routinization is that police and the judiciary may believe that the stop-level hit rates generated by a statistical model are inaccurate. However, if a model does not have the necessary information to generate accurate estimates, it is the responsibility of the police department to collect more and better data.[30] Another difficulty lies in determining whether a difference in hit rates by race is sufficiently large as to provide convincing evidence of discrimination. For example, if a SHR analysis shows that within each location, black CPW stops have a hit rate of 3%, while white CPW stops have a hit rate of 4%, is this difference evidence of discrimination? What if the white hit rate were 10%? Similarly, in a benchmark test, who will determine if differences between the racial distribution of stopped *Terry* suspects and a chosen benchmark distribution are meaningful? Furthermore, given a convincing statistical analysis suggesting discriminatory policing (e.g., large within-location racial disparities demonstrated by a rigorous SHR analysis), should the burden shift to law enforcement to provide a race-neutral justification? Although burden-shifting occurs in other contexts, like Title VII law, the Supreme Court implied in *Wayte v. United States* that

[27] *Id.* at 293.

[28] Chavez v. Ill. State Police, 251 F.3d 612, 644–45 (7th Cir. 2001).

[29] Floyd v. City of New York, 959 F. Supp. 2d 540, 603–05 (S.D.N.Y. 2013).

[30] Even given extensive, detailed, and accurate data, statistical estimates may still be inaccurate if the model itself is misspecified. However, this can be addressed in practice by using a sufficiently flexible model.

discriminatory intent requires more than simply being aware of the disparate impact resulting from a policy.[31]

The second question is, ***how should data collection and recording procedures be improved and standardized to ensure that statistical tests for discrimination are actually effective?*** The first criterion for effectiveness should be that tests give accurate results (e.g., data collection should not be skewed in some systematic way). Second, these tests should be able to directly inform policy, so data collection procedures should be prioritized if they enable tests which have immediate implications for ongoing police practice. These tests should be amenable to cross-jurisdictional comparison where possible, meaning that collection standards should be reasonably uniform across cities, counties, and states. Moreover, data collection procedures should be flexible and have enough foresight to adapt to changing circumstances. A specific example of these changing circumstances is the growing number of multiracial individuals in the United States; current data collection procedures usually only record one race for a stopped suspect. However, collecting more data introduces costs associated with storing, searching, and releasing this information (e.g., sensitive data may require substantial redaction before public release). Additionally, if more detailed information is kept on those who are stopped, and if those who are stopped are more likely to be disadvantaged or from a minority group, then burdens associated with privacy violations will fall more heavily on these minority groups.[32] Moreover, recording data free from error is difficult, although some scholarly work has introduced potential institutional and legislative solutions to minimize the occurrence of errors in criminal justice databases.[33]

Answers to these two questions will require understanding *Terry* stops as programmatic, not isolated occurrences. By thinking about stop-and-frisk as a program, courts may be more likely to allow the type of statistical evidence which requires large amounts of data. Beyond using existing corpora of data to audit police actions, a broader goal should be to conduct post-hoc evaluations of all substantive uses of technology-related policy initiatives. However, at the very least, researchers, police departments, and the judiciary should use statistical techniques like SHR analysis to produce evidence-based practices in order to make a more efficient and equitable criminal justice system.

[31] Wayte v. United States, 470 U.S. 598, 610 (1985) (quoting Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 279 (1979)).

[32] *See* Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. Pa. L. Rev. 327 (2015) (exploring risks associated with collecting large quantities of criminal justice data).

[33] *See, e.g.*, Wayne A. Logan & Andrew Guthrie Ferguson, *Policing Criminal Justice Data*, 101 Minn. L. Rev. 541, 596–611 (2016).